

A RDF-based Digital Library System

Title	A RDF-based Digital Library System
Author	Yan Han Systems Librarian University of Arizona Libraries 1510 E. University Blvd. Tucson, AZ, 85721-0055 hany@u.library.arizona.edu Tel.: (520)307-2823 Fax: (520)621-8276
Created Date	2005-08-01
Updated Date	2005-08-23

Keywords: RDF, digital library, content management systems, semantic web, interoperability, rangelands, sesame.

Abstract:

This article first introduces the needs for a true interoperability environment that allows information and its context can be transfer across domains and applications. Then it describes an approach to build a RDF-based digital library system at the University of Arizona Libraries. The system architecture consists of a storage layer, a metadata management and semantic layer, a common service layer and an application layer. The system is an artifact of the RDF model and also uses an RDF database to facilitate internal management of information resources. The article presents background for a journal delivery project and reports the implementation of the journal application using Java Servlet technology. Issues about metadata management such as various metadata formats for specific communities, and MARC to DC mapping are discussed.

1. The Needs for a True Interoperable Information Environment

Libraries, cultural heritage institutions and learning organizations manage and provide access to vast stores of information. This information is in multiple file formats and has typically been described for a specialized community's needs. Acquired for many years, this structured information would be of enormous value especially if it could be easily discovered and its meaning would be known across domains. It was noted that libraries have not exploited traditional or emerging tools for *knowledge organization* to their full potential. The report also states that research is specifically needed in how tools of knowledge representation can be used to improve users' experience in searching, navigating and comprehending digital collections (Caplan et al., 2003). *Interoperability* is essential in this environment. As interoperability moves to the foreground of research topics, the report also notes that "the digital library community lacks a commonly accepted architectural model" and "interoperability relies on standards, but many of the standards in the digital library arena are suspect in terms of their quality, applicability and durability"(Caplan et al., 2003, p14). There is a recommendation that libraries develop architectural models of standards-based, content-neutral repositories to serve as a foundation for higher level standard services (McLean and Lynch, 2004).

Digital repositories have become an essential foundation for delivering content, including scholarship, education, and cultural heritage. Some established reference models are beginning to emerge (e.g. Open Archival Information System (OAIS)). However, this content exists in specific context. Scholarship such as journal publishing and Electronic Theses and Dissertations (ETD) requires unique features for publishing, searching, discovery, and management. Education such as classroom instruction and distance learning often need their own metadata and workflows to achieve their goals. Cultural heritage programs have their own specific audiences and needs such as registration management. Therefore, all these communities tend to develop their own metadata element sets to fulfill their unique needs. Namely, Dublin Core, MARC, MODS, and METS for library community, EAD for archives, IEEE LOM and SCORM for learning community.

The current practice of encapsulating content within a well-defined purpose-based repository limits its ability to be discovered and repurposed in new and innovative ways. The traditional and current practices of exchanging information across domains and systems include Z39.50 and OAI-PMH. These practices somehow provide simple and easy means to integrate content from different sources and systems. However, the meaning of the content is described within the context of a specific community and, often, bound to a specific application. Although it is possible to locate the information stored within the repository for a new use, it was not facilitated by the repository's architecture or the system. For example, a person searching for "American Lion" will not be directly led to content that has been described with the term "Puma" which is much more common in a controlled vocabulary set. The user may stumble across this information within the course of the search, but the connection is not facilitated by the current digital library infrastructure. To leverage legacy investments and continue to support needs of specialized communities, it

is crucial to create a true interoperability environment where information not only meets the needs of specific communities, but also can be shared with others.

2. Introduction to the Journal Delivery Project

The University of Arizona (UA) has been a part of AgNIC (<http://www.agnic.org>), an initiative of the National Agricultural Library (NAL) since 1995. The mission of the UA AgNIC project is "to provide electronic access to the full scope of information in the field of rangelands and rangeland management to people everywhere and of all knowledge levels by collecting, creating, evaluating, and organizing relevant resources." The UA AgNIC project team is composed of librarians, technical staff, and rangeland experts from several collaborating units in the UA Library and College of Agriculture and Life Sciences.

In April 2004, NAL announced that funding for small cooperative agreements was available to AgNIC partners to build full-text content, which can be delivered to users through AgNIC. The library submitted a 1-year project proposal for digitizing 20 volumes (v.1-20, 1979-1998) of the Journal *Rangelands* and providing open access to the journal. The proposal was funded and a working plan was developed, including copyright clearance, digitization standards and process, metadata crosswalk, and development of a journal delivery application.

3. A RDF-based Digital Library System

We are researching on building a semantic digital library system, which can be used by libraries, cultural heritage and learning organizations, to provide robust support for specialized applications while simultaneously providing a technical foundation for meaningful access of content across the information structures of specialized heterogeneous systems. In addition, cross-domain meaning will be established by encoding organizational tools such as taxonomies, controlled vocabularies and thesauri in a machine-readable format so that essential links and crosswalks can be established. The building of a modular application on the semantic digital library system for journal delivery is a part of this research.

For the past six years, our research group has been evaluating and/or implementing several digital library systems to find a single digital library system framework to manage and provide access to our content. Like many libraries, our content is diverse and includes journal publishing and delivery, ETD, learning objects, archival materials, and museum items. The divergence of metadata standards and content formats among these specific domains and communities is necessary to address the specialized business and operational needs of these communities. It will continue to be an important attribute of the whole networked information environment for the foreseeable future. Therefore, there are valid requirements to meet special needs of these communities, but at the same time it is also crucial for users to be able to discover information across the domains and communities without having knowledge of all of the domains' specialized knowledge organization and

system structure and services.

The systems that we evaluated and/or implemented include commercial and open source systems such as DSpace <<http://www.dlearn.arizona.edu>>, Fedora, Greenstone, CWIS <<http://www.grow.arizona.edu>>, and SiteSearch <<http://jrm.library.arizona.edu>> (Han, 2004). Our systems analysis defined various metadata and systems requirements and specified that a successful system should be flexible and accommodate multiple digital formats and multiple metadata standards with related services and tools to fulfill current needs. In addition, the design must be modular and standards-based so that it can interoperate with other systems and also accommodate future needs. It has been revealed a number of shortcomings in commercial and open source digital library products (Fedora, 2004). Some products have made progress in addressing some of the issues. However, the following disadvantages were found:

- All products are narrowly focused on specific metadata standards that offer good solutions for specific purposes. Library-focused products usually use Dublin Core (DC), METS and MODS, while archival digital library systems tend to use EAD.
- Most products have predefined hierarchical structure that limits users' experience (e.g. discovery, access) and administrators' management (e.g. categorizing, extending the hierarchy).
- Products did not allow for arbitrary, multiple semantic relationships to be encoded for objects and for meaning to be shared across domains and collections.

We discovered the Resource Description Framework (RDF) <www.w3c.org/RDF/> by the W3C. RDF, a core part of the Semantic Web, uses Uniform Resource Identifiers (URIs) to uniquely identify resources and provides a framework to describe a resource in terms of its properties and its relationship with other resources. RDF uses a *graph* of nodes and arcs representing the resources, and their properties and values (W3C, 2004b). It is intended for situations in which this information needs to be processed by applications and other software agents, rather than being displayed directly to a user. It provides a common framework for expressing the information to be shared across applications. Since it is a common framework, application designers can leverage the availability of common RDF parsers and processing tools. The ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created. As an example, the Dublin Core Metadata Initiative has adopted the RDF model (W3C, 2004).

We are building a system with multiple layers. The system is composed of:

- A storage layer that is standards-based, content-neutral and metadata-extensible using RDF and RDF Schema (RDFS).
- A metadata management and semantics layer that can be used for metadata management (e.g. support multiple descriptive metadata formats and other types of metadata), ontologies, and taxonomy. Knowledge management such as transferring of meaning across domains is addressed in this layer.
- A common services layer that is standards-based and allow application specific and heterogeneous access. For example, metadata services such as OAI-PMH, linking services such as OpenURL, and search service.
- An application layer on top of common service layer that allows different

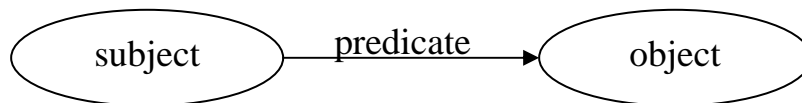
applications to fulfill diverse needs of communities.

4. Implementation

The journal delivery application is our first step to build a specific prototype application based on the digital library system and work as a concept-proof project. The storage layer is the core part of any digital library systems. Digital library systems such as DSpace, CWIS, SiteSearch and Greenstone usually have tight connection between systems' functionalities and their database schemas. Since a digital library system usually has its specific system and user requirements such as an intended metadata format, workflow, and user interface, it tends to have its own construct of multiple tables in database (e.g. Oracle, MySQL) to model complex relationship between entities.

Compared to complex relational database schema, entities in the RDF framework are simpler and more straightforward. There are only three entities in the framework: *subject*, *predicate*, and *object*. A *subject* or an *object* is a resource (a thing to be described); while a *predicate*, an aspect of a resource, expresses the relationship between a *subject* and *object*. In the RDF graph model, a subject or an object is a *node* and a predicate is represented as an *arc* in a graph. The three terms form a subject-predicate-object statement (called a *triple*, see figure 1) and consequently multiple triples form a RDF graph (W3C, 2004a.) (W3C, 2004b). The powerfulness of RDF lies in its graph model to describe the complex networked information environment.

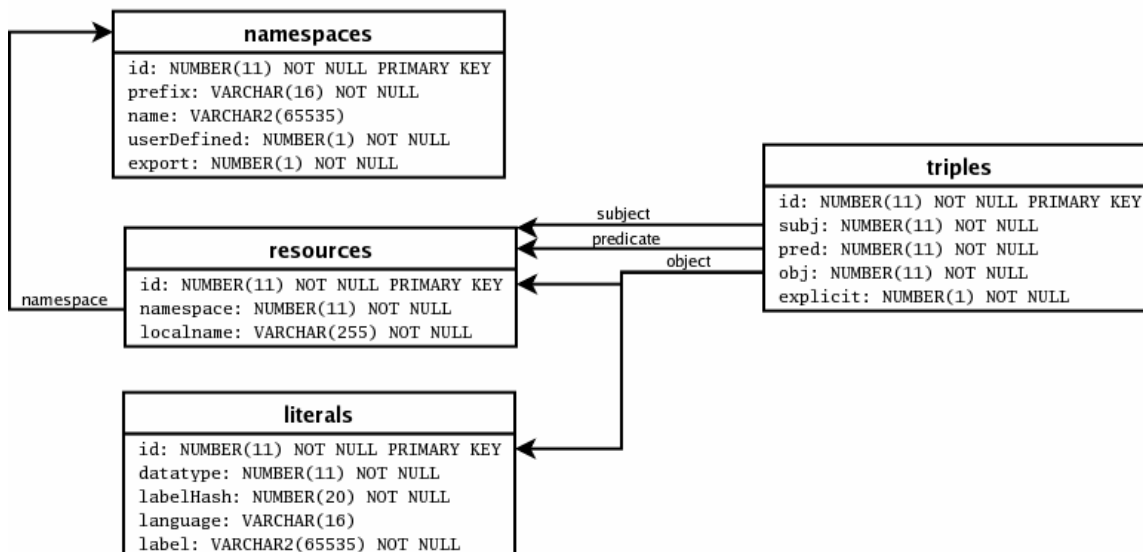
Figure 1: A RDF Triple (W3C, 2004)



We selected an open source RDF database, Sesame, as the storage layer for the semantic digital library system due to its popularity, support and maturity. Sesame is an open source RDF database with support for RDF querying and inferencing. With support of several RDF query languages, it has flexible access API and good support (Broekstra, 2002). Building on top of Sesame RDF database, our database schema has extremely simple structure (see figure 2), as it has only 4 tables: resources, triples, namespaces, and literals. Table “resources” stores subjects and objects; while table “triples” contains subject-predicate-object statements (triples). Consequently, the simple database schema also reduces programming efforts whenever there are needs to connect database to perform queries. Database management issues are also reduced.

Figure 2: Database Schema for the RDF-based Digital Library

Sesame Database Schema



The metadata management layer might be the most complicated layer in terms of the framework, design, and programming. To be a generic semantic digital library, it must be able to provide supports for multiple applications (e.g. the journal delivery application) by handling multiple descriptive metadata formats such as DC, MARC, and EAD. On the other hand, it may need to deal with other types of metadata such as technical metadata and preservation metadata. In addition, knowledge management requires complicated ontologies, a variety of controlled vocabularies, and multiple taxonomies to work together.

At this moment, we do not have a completed product to support multiple metadata formats. Only qualified DC was implemented to handle all the needs of the journal delivery application. 1220 MARC records were received from NAL, each of them corresponding to one article in the *Journal of Rangelands*. Since the current metadata management layer is using qualified DC, a simple crosswalk (mapping) is necessary for transferring records in MARC to their counterparts in DC. MarcEdit contains a XSLT script that provides MARC to DC mapping based on MARC to DC crosswalks published by the Library of Congress. It is worth noting that “not all possible MARC fields are included in this mapping, but only those considered useful for broad cross-domain resource discovery”(Library of Congress, 2001). The original script misses some important MARC fields, including 246 (alternative title), 300 (physical description, 651 (subject), and 773 (host item entry). These fields are significantly important for the collection and at the same time improve customers’ experience in browse, search and discovery. The XSLT script was modified to meet our needs for mapping all related MARC elements.

In the common services layer, modules have been written to enable the current OAI-PMH metadata service and OpenURL linking service. These services will be common services not only for the journal delivery application, but also for future applications.

The journal delivery application lives in the application layer. Based on systems analysis,

the application must provide basic browse and search functions. For example, it must provide hierarchical browse feature and search function for metadata fields such as author, title, and keyword. In addition, its user interface should be designed with information architecture in mind. The application is using Java Servlet and Java Server Pages (JSP) technologies with Model-View-Controller (MVC) design pattern in mind. Since journals usually have a pre-defined hierarchical structure, the following classes are constructed in hierarchy: article, issue, volume, and journal. Corresponding helper classes are also designed to handle other interactions with the RDF database, and display.

The application has been running since April 2005. With respect to our systems requirements, we believe that the journal delivery application achieve its functionality. Furthermore, the application and its underlying architecture provide a solid basis for integrating other applications for their specific needs. The next step is to build an ETD application, which will handle ETD needs and be in the application layer by using the same storage and metadata management architecture.

Summary

Starting with presenting the needs for a true interoperable networked information environment, this paper describes some of current issues in current digital library systems and the lack of true interoperability across domains. The paper reports that the UA library's approach to build a semantic, standards-based, and content-neutral digital library system. It is hoped that the semantic digital library system will be able to allow better interoperability. Extensive research has been carried out to investigate an appropriate framework to represent information resources in the system by using the RDF graph model. The system utilizes the RDF database as the core storage layer and allows upper application layer to build specific applications to meet various needs. Experience shows that the use of the RDF database results in a very simple database schema and reduces programming efforts involving the journal delivery application.

Acknowledgement

This paper is based on a proof-of-concept project at the University of Arizona Libraries. I would like to thank my colleagues' work and inputs for the project. Krisellen Maloney and Travis Bowen are very active in the project. We have been working together to design the system architecture, develop systems requirements and implement the system.

References

Broekstra, J., Kampman, A., & Harmele, F., (2002), *Sesame: A Generic Architecture for Storing and Querying RDF*, the International Semantic Web Conference 2002, Sardinia, Italy. Available at: <http://www.openrdf.org/doc/papers/Sesame-ISWC2002.pdf> (accessed 10 Aug 2005).

Caplan, P., Barnett, B., Bishoff, L., Borgman, C., Hamma, K., & Lynch, C. (2003), *The report of the workshop on opportunities for research on the creation, management, preservation and use of digital content*. IMLS. Available at: <http://www.imls.gov/pubs/pdf/digitalopp.pdf> (accessed 1 Oct 2004)

Fedora. (2002), *Mellon Fedora Technical Specification* University of Virginia Library, Cornell University. Available at: <http://www.fedora.info/documents/master-spec-12.20.02.pdf>

Fedora. (2004), *Fedora, phase 2*. Available at: http://www.fedora.info/documents/fedora2_final_public.html (accessed 1 Oct 2004)

Han, Y. (2004), "Digital content management: The search for a content management system". *Library Hi Tech*, 22(4), 355-365.

Maloney, K., Han, Y., & Bowen, T. (2005), *A semantic RDF-based Digital Library System*. A Proposal submitted to IMLS National Leadership Grants.

McLean, N., & Lynch, C. (2004), *Interoperability between Library Information Services and Learning Environments – Bridging the Gaps: A Joint White Paper on Behalf of the IMS Global Learning Consortium and the Coalition for Networked Information*. IMS Global Learning Consortium and CNI. Available at: http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf

NASA - Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: NASA. Available at: <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>

W3C. (2004). *RDF Primer*. Available at: <http://www.w3.org/TR/rdf-primer/#intro> (access 1 Oct 2004).

W3C. (2004), *Resource Description Framework (RDF): Concepts and Abstract Syntax*. Available at: <http://www.w3.org/TR/rdf-concepts/> (access 1 Aug 2004)

W3C. (2004), *Resource description framework (RDF): Dublin core metadata initiative*. Available at: <http://www.w3.org/RDF/> (access 1 Aug 2004).